Scoring methods for joint damage on plain radiographs in rheumatoid arthritis : progressive understanding of methodological issues

Summary

The disease rheumatoid arthritis (RA) is characterized by a chronic (sterile) inflammation of the joints, mainly localized in the smaller joints of the hands and feet. Persistent arthritis due to RA often leads to permanent damage of the cartilage and underlying bone. The damage can be visualized on plain radiographs. Anti-inflammatory drug can inhibit the inflammation and may slow down or prevent deterioration or occurrence of joint damage. In order to investigate whether these drugs indeed are able to moderate joint damage, plain radiographs are included in clinical research. Further, to be able to analyse the efficacy of the drug under investigation, the amount of damage on cartilage and bone needs to be quantified. Quantification of joint damage due to RA is called "scoring" and the standardized approach of scoring pre-selected joints is called a "scoring method". How well the ability of a drug to modify joint damage can be investigated mainly depends on the validity, reproducibility and responsiveness of the scoring method used. Several radiological scoring methods have been described and evaluated in the past. The methods developed by Sharp and Larsen and their modifications – like the Sharp/van der Heijde method and the Larsen/Scott method- are the most widely used methods at this moment.

Chapter 2-9 of this thesis describe research performed to solve several outstanding issues regarding scoring joint damage by means of the Sharp/van der Heijde method and the Larsen/Scott scoring method. This with the goal to use the optimal radiological scoring method in future clinical research on RA and consequently to be able to answer research questions with less patients and/or less money.

Chapter 2 describes how the discriminative capacity and responsiveness of the Sharp/van der Heijde and the Larsen/Scott relate to each other. For this purpose so-called generalizability analyses were used. The analyses showed that the discriminative capacity and responsiveness of the methods were both highly acceptable and comparable if mean group scores were calculated and compared. If the percentage of patients that actually showed deterioration were calculated and compared, the Sharp/van der Heijde scores were measured with more precision.
With these analyses also the issue of the optimal number of investigators (scorers) used in clinical research to score joint damage could be addressed. The analyses showed that the discriminative capacity and responsiveness of both methods changed most if two instead of one scorer were used.

Prevention of erosions (destruction of the bony surface underlying the cartilage) in undamaged joints is often considered more important than prevention of

deterioration in already damaged joints. Chapter 3 however showed that important information on the efficacy of the drug investigated might be missed and/or more patients are needed if the investigation is limited to undamaged joints. It thus seems preferable to judge both undamaged as well as already damaged joints in clinical research.

In chapter 4 the so-called minimal clinically important difference (MCID) was assessed for both scoring methods. An international expert panel was used to define this MCID. The panel consisted of five experienced rheumatologists who are used to judge radiographs of patients themselves but who were not trained to quantify joint damage in a standardized way. The rheumatologists judged independently sets of radiographs that had been obtained with 1-year intervals. They were asked to state whether they saw progression of joint damage or not. If they stated that they noticed progression of damage they were further asked whether they judged that amount of progression of that amount that they wanted to change the patient's therapy. Or in other words, whether they found the amount of progression clinically relevant. The panel members were asked to judge the progression noticed for four different types of patients; patients with 1) recently diagnosed RA with mild symptoms, 2) recently diagnosed RA with severe symptoms, 3) advanced RA with mild symptoms and 4) advanced RA with severe complaints. The panel's judgments were compared with the average changes in joint damages measured by the Sharp/van der Heijde and the Larsen/Scott method. These scores were independently obtained by two experienced scorers for each scoring method.
The number of patients that actually show clinically important deterioration (or repair) of joint damage is often calculated because it can give additional information on the efficacy of the drug investigated. If one wants to calculate the number of patients that show clinically important deterioration (or repair) of joint damage one needs to know the MCID. In the past, the so-called smallest detectable difference has - as alternative- been used as cutoff value to express the continuous scores as an outcome with two answer possibilities (progression yes or no) [the smallest detectable difference is the difference between two scores that is greater than the measurement error of the scoring method. This statistical method assesses the measurement error of the scoring method that occurs while repeating the scoring]. In chapter 4, the MCID was therefore compared with this smallest detectable difference.
The results in chapter 4 showed that the panel on average judged a change in joint damage in the order of the magnitude of the smallest detectable difference of the Sharp/van der Heijde method as clinically relevant. The difference between the MCID and the smallest detectable difference of the Larsen/Scott method was remarkable: the panel judged changes smaller than the measurement error of the Larsen/Scott method already as clinically relevant. Additional analyses showed that this probably can be explained by the fact that the panel also took the narrowing of joint spaces (due to cartilage loss) into account. If one quantifies joint damage with the Larsen/Scott method, losses of cartilage are hardly taken into account in the scores.

In order to validate the judgments of the panel, the radiographs were also judged by a couple of radiologists who were highly experienced in judging joint damage due to RA. The results described in chapter 5, showed that the radiologists were more reserved in judging changes as important than the panel of rheumatologists. The panel of rheumatologists were inclined to change therapy in cases of progression of joint damage that the radiologists reported as not substantial. The radiologists further judged less sets of radiographs as progressive. This study confirmed former research that showed that the profession of the panel members can strongly influence panel results. The panel used in the study of chapter 4 could however not include radiologists. After all, this panel had to judge whether they wanted to change the treatment strategy of a patient based on the deterioration of joint damage. Radiologists are not trained in making inferences for patients' treatment strategies. In conclusion, if possible, it seems preferable to include clinicians of different profession when using an expert panel in your research, in order to be able to generalize the results as much as possible.

Research has shown that significant more progression is detected if the scorer knows the chronological order of the radiographs he/she has to judge. Explanation for this is most likely that one is able to correct maximally for variation on quality of the radiographs or variation in position of the hands and/or feet if the correct order is known. However, question remained whether knowledge regarding the correct order of the radiographs not primarily leads to overestimation of progression instead of a more precise detection. In order to answer this question, the influence of knowledge of the chronological order on the detection of clinical relevant progression of joint damage by the scoring methods was investigated in chapter 6. The clinical relevance for recent diagnosed RA patients with severe symptoms was used for this purpose and was defined by the same panel as used in chapter 4. The results showed that the scoring methods were more sensitive in detecting clinical relevant progressions if the scorers knew the correct order of the radiographs. The results showed however also that the gain in sensitivity lead to some not ignorable loss in specificity, resulting in the wrongly classification of progression of joint damage as clinical relevant in some cases.
Besides that knowledge of the chronological order leads to more sensitive scoring methods, this knowledge on the other hand also appears to make the interpretation of repair of damage difficult. Recent research has suggested that joint damage due to RA is not that permanent as assumed. In order to be able to investigate this, scorings methods should be used in which the agreement between the positive change scores (progression) is equal to the agreement between the negative change scores (repair of damage). Because this is not the fact if the chronological order is known to the scorers, this method seems not to be recommended for research at this moment.
In the course of the above described research, it became clear that the formulas used to calculate the smallest detectable difference should be adapted if

radiographs are scored in pairs, as is usual for the judgment of joint damage due to RA. With paired scores, sets of radiographs of one patient are judged simultaneously, causing the measurement error of the different scores of that patient to be related. The correct formulas to calculate the smallest detectable difference if radiographs are judged in pairs is presented in chapter 7 (called the smallest detectable change). This chapter further showed that with the original smallest detectable difference, the measurement error of detecting a change in scores is overestimated if radiographs are scored in pairs.

Research at new treatments often includes patients with a high risk at the disease outcome to be prevented/diminished (like joint damage). This is done with the aim to make the contrast in the number of patients with the eventual outcome to be prevented as large as possible between the patients in the treatment and the control group. The presence of joint damage at the start of a clinical study is one of the best predictors for progression of joint damage during the trial. Therefore, it was expected that less patients would be needed to answer a research question when patients with much joint damage would be selected for inclusion. Chapter 8 describes why this is not always true for outcomes with two answer possibilities. In summary, the effect of a selective inclusion of patients with a high risk depends on how the treatment reduces the risk on the outcome investigated. Not in all cases this is leading to the expected enhanced effect.

Chapter 9 presents the results of a study that investigated whether it is possible to detect change in joint damage on plain radiographs taken with a 3-month interval. In studies testing a new drug, the modifying effect of the new treatment on joint damage is usually investigated after 12 months. Recent research has however shown that also after six months significant differences in progression of joint damage can be found. The clinical studies that evaluate the most optimal dosage and treatment duration, so-called phase 2 studies, however usually last three months. Therefore it is interesting to investigate whether it is possible to detect progression of joint damage already in a 3-month interval. Chapter 9 showed that the Sharp/van der Heijde method was able to detect change in joint damage on plain radiographs taken with a 3-month interval, with or without known chronological order of the radiographs. Additional analyses further showed that also with a 3-month interval, the number of patients needed to detect significant differences between a treatment and control group could be acceptable when plain radiographs are taken with a 3-month interval.

Samenvatting

De ziekte Reumatoïde Artritis (RA) wordt gekarakteriseerd door een chronische (steriele) ontsteking van de gewrichten die zich met name manifesteert in de gewrichten van de handen en de voeten. De chronische gewrichtsontsteking bij RA kan leiden tot onherstelbare schade van het kraakbeen en (of) het daaronder liggend bot. Deze beschadigingen kunnen worden gezien op röntgenfoto's. Met geneesmiddelen kan de ontsteking in de gewrichten geremd worden waardoor de (toename) van de gewrichtsschade (gedeeltelijk) voorkomen of verminderd kan worden. Om tijdens klinisch wetenschappelijk onderzoek te kunnen beoordelen of een geneesmiddel deze modificerende werking heeft worden er röntgenfoto's gemaakt. De mate van schade aan het bot en kraakbeen op de foto's moet in getal uitgedrukt worden om de werking van verschillende geneesmiddelen daadwerkelijk te kunnen beoordelen en/of vergelijken. Het toekennen van een waarde aan de schade wordt 'scoren' genoemd en het op gestandaardiseerde wijze scoren van tevoren geselecteerde gewrichten wordt scoren volgens een bepaalde 'scoringsmethode' genoemd. Hoe goed een geneesmiddel kan worden beoordeeld hangt in belangrijke mate af van hoe betrouwbaar de scoringsmethode is. Dit betreft de reproduceerbaarheid (onder dezelfde omstandigheden dezelfde score geven) en gevoeligheid voor verandering van de gebruikte scoringsmethode. Verscheidene scoringsmethodes zijn in het verleden beschreven en geëvalueerd. De heden meest gebruikte methodes zijn de door Sharp en de door Larsen ontwikkelde of daarvan afgeleide methodes (zoals de zogenaamde Sharp/van der Heijde en de Larsen/Scot methode).

In dit proefschrift wordt in hoofdstuk 2-9 onderzoek beschreven naar een aantal nog openstaande vragen bij het scoren van gewrichtsschade met behulp van de Sharp/van der Heijde en Larsen/Scott scoringsmethode. Dit met als doel om in toekomstig wetenschappelijk onderzoek de optimale scoringsmethode te kunnen gebruiken waardoor met minder patiënten en minder geld een betrouwbaar antwoord kan worden verkregen.

Hoofdstuk 2 beschrijft hoe het onderscheidend vermogen en de gevoeligheid voor verandering van de Sharp/van der Heijde en de Larsen/Scott methode zich ten opzichte van elkaar verhouden. Dit is onderzocht met zogenaamde 'geneneraliseerbaarheidanalyses'. Deze analyses lieten zien dat het onderscheidend vermogen en de gevoeligheid voor verandering voor beide methodes acceptabel tot goed en vergelijkbaar waren indien gemiddelde scores van groepen met elkaar worden vergeleken in een onderzoek. Indien echter in een onderzoek de percentages patiënten dat werkelijk toename (of afname) van schade laten zien worden vergeleken dan lijken de scores van de Sharp/van der Heijde methode met grotere precisie te worden gemeten.

Met behulp van deze analyses kon ook een uitspraak worden gedaan over het optimale aantal te gebruiken beoordelaars bij klinisch wetenschappelijk onderzoek. Het onderscheidend vermogen en de gevoeligheid voor verandering van beide scoringsmethodes bleek het meest te verbeteren indien twee in plaats van één beoordelaar werden gebruikt.

Vaak wordt in onderzoek het voorkómen van erosies (beschadiging van het oppervlak van het bot aangrenzend aan het kraakbeen) in niet beschadigde gewrichten belangrijker geacht dan het voorkómen van progressie van de schade van gewrichten die reeds erosies hebben. Hoofdstuk 3 laat echter met behulp van gegevens van een recent geneesmiddelenonderzoek zien dat belangrijke informatie over het geneesmiddel kan worden gemist en/of meer patiënten nodig zijn indien alleen de onbeschadigde gewrichten worden beoordeeld Het lijkt dus raadzaam om in wetenschappelijk onderzoek de gewrichtsschade in zowel de onbeschadigde als in de reeds beschadigde gewrichten te beoordelen, zodat mogelijk met minder patiënten een antwoord kan worden verkregen op de vraagstellingen.

In hoofdstuk 4 is voor beide scoringsmethodes het *kleinste klinisch relevante verschil* in gewrichtsschade geschat. Hiervoor is gebruik gemaakt van een internationaal expertpanel. Dit panel bestond uit vijf ervaren reumatologen die in de praktijk zelf hun röntgenfoto's beoordelen maar niet getraind waren om op een gestandaardiseerde wijze een getal aan de schade te geven. De reumatologen werden gevraagd om onafhankelijk van elkaar fotosets met één jaar tussentijd te beoordelen op het wel of niet ontstaan en/of toenemen van de gewrichtsschade. Indien ze progressie van de gewrichtsschade zagen, moesten ze aangeven of ze de verandering van dusdanige aard vonden dat ze de therapie van de patiënt wensten aan te passen. Met andere woorden, aangeven of de progressie wel of niet gezien moest worden als klinisch relevant. Dit werd gevraagd voor vier verschillende typen patiënten; patiënten met 1) recent vastgestelde RA en milde klachten, 2) recent vastgestelde RA en ernstige klachten, 3) langer bestaande RA en milde klachten en 4) langer bestaande RA en ernstige klachten. Het oordeel van het panel werd vervolgens vergeleken met de gemiddelde Sharp/van der Heijde en Larsen/Scott progressiescores (het verschil tussen de score bij het begin en één jaar later) van twee beoordelaars. Hiervoor werden per scoringsmethode de röntgenfoto's onafhankelijk gescoord door twee verschillende ervaren beoordelaars.
Het aantal patiënten dat daadwerkelijk een klinisch relevante verandering laat zien, kan interessante aanvullende informatie opleveren bij het analyseren van klinisch wetenschappelijk onderzoek. Om dit te kunnen doen moet men echter weten wat het kleinste klinisch relevante verschil is. In het verleden is als alternatief vaak het zogenaamde *kleinst meetbare verschil* gebruikt als afkappunt om de continue scores van de scoringsmethodes uit te drukken als een maat met twee uitkomsten (wel of geen progressie) [het kleinst meetbare verschil is een verschil dat groter is dan de meetfout van de scoringsmethode. Deze statistische maat schat de meetfout van de scoringsmethode die optreedt

tijdens het reproduceren van scores). In hoofdstuk 4 is dientengevolge het door het panel bepaalde kleinst klinisch relevante verschil vergeleken met het kleinst meetbare verschil.
Hoofdstuk 4 liet zien dat het panel gemiddeld gezien schade ter grootte van het kleinst meetbare verschil van de Sharp/van der Heijde scoringsmethode reeds als klinisch relevant beoordeelde. Het verschil tussen het kleinst meetbare verschil van de Larsen/Scott methode en het kleinste klinisch relevante verschil was echter opmerkelijk. Bij progressie kleiner dan de meetfout van de Larsen/Scott methode oordeelde het panel dat ze de therapie al wilde wijzigen. Aanvullende analyses lieten zien dat dit waarschijnlijk grotendeels verklaard kan worden door het feit dat het panel ook in belangrijke mate de gewrichtsspleetvernauwing (een gevolg van kraakbeenverlies) in hun oordeel meenam. De gewrichtsspleetvernauwing wordt bij de Larsen/Scott scoringsmethode slechts in geringe mate meegenomen bij het scoren van de gewrichtsschade.

Ter validatie van het in hoofdstuk 4 gebruikte expertpanel, werden de röntgenfoto's ook voorgelegd aan twee radiologen met uitgebreide ervaring in het beoordelen van gewrichtsschade tengevolge van RA. De resultaten in hoofdstuk 5 lieten zien dat de radiologen meer gereserveerd waren in hun oordeel dan het panel reumatologen. Het panel reumatologen gaf aan de therapie te willen aanpassen bij veranderingen in radiologische schade die door de radiologen niet als substantieel beoordeeld werden. Bovendien beoordeelde de radiologen in totaal minder fotosets als progressief. Deze studie bevestigt de in het verleden uitgevoerde onderzoeken die lieten zien dat het beroep van een panellid de uitslagen van een panel sterk kunnen beoordelen. Bij het experiment uitgevoerd in hoofdstuk 4 was het echter niet mogelijk om ook radiologen te includeren Dit panel moest immers beoordelen of ze de therapie van de patiënten wilde wijzigen en radiologen zijn niet getraind in het doen van uitspraken over het therapeutische beleid van een patiënt. Samenvattend lijkt het dus echter wel raadzaam om leden van verschillende beroepen te laten deel nemen in een expertpanel indien dit mogelijk is. Dit met het oog op een zo groot mogelijke generaliseerbaarheid van de resultaten

Uit onderzoek is gebleken dat indien de chronologische volgorde van een serie röntgenfoto's bekend is bij de beoordelaar(s) er significant meer progressie gemeten wordt dan wanneer onbekend is welke foto de eerste in de tijd is. Een logische verklaring voor deze bevinding is dat indien de volgorde van röntgenfoto's bekend is maximaal gecorrigeerd kan worden voor variatie in kwaliteit of in de positie van de handen/voeten tussen de foto's. De vraag bleef echter of de kennis over de correcte volgorde van röntgenfoto's niet juist leidt tot een overschatting in plaats tot een preciezere schatting. Om hierover een uitspraak te kunnen doen, is in hoofdstuk 6 de invloed van het al dan niet bekend zijn van de chronologische volgorde van een serie röntgenfoto's op het detecteren van klinisch relevante progressie door de scoringsmethodes onderzocht. Hiervoor is gebruik gemaakt van het oordeel van hetzelfde

expertpanel als beschreven in hoofdstuk 4 voor de groep RA patiënten met ernstige klachten en waarbij recent de diagnoses RA is gesteld. De resultaten lieten zien dat de scoringsmethodes veel gevoeliger waren voor het oppikken van patiënten met klinisch relevante progressie indien de volgorde van de röntgenfoto's bij de beoordelaars bekend was. Dit ging echter wel ten koste van enige specificiteit van de methodes, resulterend in het soms ten onrechte classificeren van de progressie als klinisch relevant.
Naast het feit dat scoren met kennis over de volgorde van de röntgenfoto's blijkt te leiden tot meer gevoelige scoringsmethodes, blijkt het geven van deze kennis echter ook de interpretatie van herstel van schade te bemoeilijken. Recent onderzoek suggereert dat de gewrichtsschade bij RA toch niet zo definitief is als men gedacht heeft. Om dit beter te kunnen onderzoeken, heeft men scoringsmethodes nodig waarbij de overeenstemming tussen de beoordelaars bij zowel progressie als bij herstel van de gewrichtsschade gelijk is. Daar dit niet het geval is indien de chronologische volgorde bekend is bij de beoordelaar(s), lijkt deze methode momenteel niet aan te raden voor het doen van wetenschappelijk onderzoek.

Gedurende bovenstaand onderzoek werd duidelijk dat de formules gebruikt om het kleinst meetbare verschil te bepalen aangepast moeten worden indien men dit toepast op gepaarde waarnemingen, zoals gebruikelijk bij het beoordelen van de gewrichtsschade bij RA. Met gepaarde waarnemingen word hier het tegelijk beoordelen van opeenvolgende röntgenfoto's van één patiënt bedoeld, waardoor de meetfout van de verschillende scores van één patiënt met elkaar samenhangt. In hoofdstuk 7 wordt de juiste formule gepresenteerd om het kleinst meetbare verschil te bepalen indien de scores 'gepaarde' waarnemingen zijn. Tevens laat dit hoofdstuk zien dat met de originele formules de meetfout wordt overschat indien scores 'gepaard' zijn bepaald.

Bij onderzoek naar een nieuwe behandeling worden vaak patiënten geïncludeerd die een hoog risico hebben op de door de nieuwe behandeling te voorkomen/verminderen van de ziektemaat (zoals gewrichtsschade). Dit wordt gedaan om het te verwachten verschil in patiënten tussen de behandelgroep en de controlegroep zo groot mogelijk te maken. Hierdoor kan met minder patiënten reeds een goed oordeel worden geveld over de nieuwe behandeling. Daar het hebben van gewrichtsschade één van de sterkste voorspellers is voor het ontstaan van nieuwe gewrichtsschade, zou men verwachten dat bij selectie van patiënten met veel gewrichtsschade met minder patiënten een verschil tussen de behandelgroepen kan worden aangetoond. In hoofdstuk 8 wordt beschreven waarom dit bij uitkomstmaten met twee antwoordmogelijkheden (zoals wel of geen progressie van gewrichtsschade) echter niet altijd opgaat. Samenvattend blijkt het effect van selecteren van een 'hoog risicogroep' af te hangen van de wijze waarop een behandeling de kans op de te onderzoeken uitkomstmaat vermindert. Niet in alle gevallen levert dit het verwachte gunstige effect.

In hoofdstuk 9 worden de resultaten beschreven van een studie die onderzocht of met röntgenfoto's ook veranderingen in gewrichtsschade kunnen worden aangetoond in een periode van drie maanden. Volgens de richtlijnen van de Amerikaanse FDA (Food & Drug Association) wordt in de meeste onderzoeken naar nieuwe geneesmiddelen voor RA pas na 12 maanden geëvalueerd of het nieuwe middel een modificerende werking heeft op de gewrichtsschade. Recent geneesmiddelenonderzoek heeft echter aangetoond dat ook na zes maanden reeds verschillen in gewrichtsschade kon worden gezien. Daar de klinisch wetenschappelijk studies die de optimale dosis en behandelingsduur van een geneesmiddel onderzoeken, de zogenaamde fase 2 studies, meestal drie maanden duren, is het dus interessant om te onderzoeken of je met röntgenfoto's reeds na drie maanden verschillen in groepen kan detecteren. Het onderzoek beschreven in hoofdstuk 9, liet zien dat de Sharp/van der Heijde methode significante veranderingen kon detecteren met een tussentijd van drie maanden. Aanvullende analyses lieten bovendien zien dat ook bij een interval van drie maanden, onder de juiste omstandigheden, het aantal patiënten dat nodig zal zijn om verschillen tussen een behandel- en controlegroep aan te tonen acceptabel is.